DOCUMENT RESUME

ED 180 295 HE 01? 959

AUTHOR McIntyre, Charles J.
TITLE Peer Evaluation of Teaching.
PUB DATE Aug 78
NOTE 7p.: Paper presented at the American Psychological Association Convention (Toronto, Canada, August 31-September 1, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Faculty: *College Instruction: Effective Teaching: *Evaluation Methods: *Faculty Evaluation: Higher Education: *Peer Evaluation: Student Evaluation of Teacher Performance: Teaching Skills

ABSTRACT

The importance and issues of peer evaluation of teaching are discussed. It is suggested that a more tangible relationship between professional advancement and teaching effectiveness is necessary to motivate faculty to improve instruction, and that evaluation can provide essential information to a teacher to help improve teaching quality. Problems and limitations in student evaluation of teaching are explored, including student ability to adequately judge the propriety of the emphasis or de-emphasis of various elements of the course content and the adequacy of the examination system. Problems in faculty peer evaluation, such as time requirements to make reliable judgments and faculty qualifications to make judgments about other specialities (even within the same department), are examined. The need to develop procedures for evaluation is considered as is the option of ignoring teaching in the evaluation process. It is concluded that more experimentation and development are necessary to arrive at a means of involving faculty in a fair and effective manner. (PHR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED180295

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Charles J. McIntyre
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PEER EVALUATION OF TEACHING[1]

Charles J. McIntyre
Director
Office of Instructional Resources
University of Illinois at Urbana-Champaign

To the extent that it occurs at all, the evaluation of one's colleagues' teaching appears to be an unsystematic process in most institutions. Not very much seems to be known about how to go about systematically evaluating the teaching of one's peers. The literature is very thin and generally unsatisfactory.

In this paper I shall try to suggest why it is a topic which should interest us more than it appears to, and I shall try to indicate what some of the significant issues are.

This decade has brought with it a substantial apparent increase in the interest being expressed by colleges and universities in the improvement of instruction and in the evaluation of teaching. This apparent increase in interest is evidenced by institutions establishing programs devoted to helping faculty improve their teaching, by the formation of instructional development interest groups in professional organizations and by the formation of consortia focused upon faculty development.

Instructional improvement and evaluation are, of course, related activities, although they need not always have a common stimulus. I believe that the recent surge of interest in improving instruction derives from the intereaction of consumerism among students with the predicted decline in the number of students wishing to go to college. Consumerism teaches that people ought to have the opportunity to know what it is they are paying for and to then have sufficient information to decide whether they have gotten it. The predicted decline in enrollment seems to be promoting competition for students among colleges and universities. A reputation for good teaching may be important for success in that competition.

Evaluation of instruction enters the picture in several ways: First, if an institution wants to help motivate faculty to improve its instruction, it seems that there ought to be some tangible relationship between professional advancement at that institution and teaching effectiveness; secondly, even if neither the administration nor the faculty at an institution is interested in giving more than lip service to the idea of rewarding good teaching, it is quite possible that Federal affirmative action guidelines may require that there be a definable relationship between advancement at the institution and what the

---

[1]Paper prepared for presentation at the American Psychological Association Convention, Toronto, Canada, August 31 -- September 1, 1978.

HE011 959

institution says is an important part of a faculty member's job—teaching; and thirdly, evaluation can provide essential information to a teacher to help improve his or her teaching.

The first two applications of evaluation, mentioned above, are primarily for administrative purposes and generally go under the rubric of "summative evaluation." The third application is intended to be directly helpful to the instructor in improving instruction and is usually described as "formative." Parenthetically, let it be noted here that one of the unanswered questions of faculty evaluation is whether both summative and formative evaluation can be carried out optimally by the same people, with the same instruments, and with the same process. Efficiency certainly indicates that it be done that way, but there is obviously the danger that one will interfere with the other. For instance in summative evaluation, one would expect the person being evaluated to want to put his best foot forward, but for formative evaluation one may want to learn the worst, whatever that may be.

In any case, if the institution believes that some kind of systematic evaluation of teaching is necessary, then one must ask what will be the nature of that evaluation? Very likely most evaluation of teaching is carried out by colleagues and administrators in an impressionistic, unsystematic way. But if we want to consider, as I do here, only formal, systematic evaluation, then the method of choice for most institutions at present is some form of ratings by students. The use of student ratings is usually accompanied, I believe, by a fair amount of grumbling, skepticism, and viewing with alarm, but nevertheless has obtained such popularity as it has, I believe, because they are relatively easy to administer, they can be constructed reliably, they have at least face validity, and they generate lots of data which satisfy our passion for numbers that can be manipulated, organized, ordered, normed and studied. It is not my purpose to argue with any of these advantages of student ratings. I am personally convinced that student ratings are an important component of an evaluation system. I only object to the fact that they are usually the only systematic formal component of the procedure for evaluating teaching. I acknowledge that peers very probably have a general idea about each other's teaching, based upon impressions they have gotten of each other in faculty meetings, seminars, social situations, and their knowledge of student opinion. This general impression surely gets factored into recommendations that they make about each other, but it is not normally systematic, explicit or formal.

Both faculty and evaluators are uneasy about leaving the entire systematic evaluation of teaching up to the students only. They are concerned that occasionally students may be unduly bedazzled by a showman or irritated by a stiff task master. They are also properly concerned over the probability that students may not be able to adequately judge all the factors that constitute good teaching. For instance, can students judge:

1. Appropriateness of materials being taught, within the curriculum of which it is a part?

2. The currency of scholarship which supports the course being evaluated?

3. The propriety of the emphasis or de-emphasis upon various elements of the course content?

4. The adequacy of instructional materials used in comparison with what is actually available?

5. The adequacy of the examination system?

Others doubtlessly would want to add to or amend this list, but it is sufficient for my purpose.

One should begin to get a sense of why the literature is so sparse in this area. It is probably because peers do not evaluate each other in these terms often enough or systematically enough to permit much study of the process. There are severe practical and psychometric problems in making, collecting and interpreting judgments by colleagues of each others' teaching. Some of the practical problems are these:

1. Where will enough faculty find enough time to visit classrooms and examine materials with sufficient care to arrive at reliable judgments? I suspect time and effort could be invested if—but only if—faculty were persuaded that that is worth more to themselves and to their institution than something else which would have to be cut back.

2. Can essential collegiality in a department be maintained when colleagues are looking over each other's shoulders as closely as a peer evaluation system would seem to require? At present colleagues maintain reasonably good relationships with each other and still make discriminating judgments when they evaluate each other's research. There is a difference, however, in that the research is usually evaluated in the first place by referees and editors of journals and other persons not within the institution. Hence, the assessment being made is concerned more with what others outside of the institution think of the colleague's work than with the work itself. There is also some interesting evidence which suggests that colleagues tend to rate each others' teaching more leniently than students do, which is the reverse of what most critics of student evaluation would apparently expect. Perhaps this means, where those results have been found, that faculty solve the collegial problem simply by being generally nice to each other. They do unto others as they hope others will do unto them, but this yields rather uniform ratings which aren't much use in making discriminations.

3. Do faculty, even in the same department, really know enough about each other's specialities to evaluate each other's courses? Better than students, certainly, but probably not all faculty are equally well-qualified to evaluate all other faculty. Therefore, one colleague's ratings are not necessarily as appropriate as another's.

The psychometric problems are also substantial. The number of raters is usually small, the amount of instruction given to raters to assure that they are rating the same things is usually minimal and the raters are often quite unevenly qualified to make their evaluations. Therefore, one would expect that the reliabilities of the ratings will be quite low. This indeed is usually what is found. Where reliability is low, whither goes validity?

Having come to validity, it is reasonable to ask "If a procedure purports to measure teaching effectiveness, does it?" The question is deceptively simple. It implies that some measurable, independent, generally agreed upon outside criteria of teaching effectiveness exist where in fact they do not. It is probably quite impossible to get academics to agree on a general formulation of what constitutes teaching effectiveness and possibly there is no reason why one should expect that, for what is applicable in one situation need not be at all applicable in another. Even so, some will expect evaluators to demonstrate validity in some straightforward way and the whole activity is severely embarrassed by the lack of criteria.

Suppose it were assumed that effective teaching is whatever it is that brings about student learning and that valid measures of teaching effectiveness ought to be correlated with differences between pre- and post-test scores. There is some research which has tried to do this, usually using student ratings (rather than peer ratings) as the other variable. The results have been inconclusive. Sometimes the correlations have been positive, sometimes not sufficiently different from zero, and in a few cases negative. Doyle and Crichton(5) discuss the problem of using student learning as a criterion. To summarize their discussion, they believe that ratings and learning may yield low correlations because of a deficiency in the ratings, a deficiency in the learning measure, a deficiency in the logic of or the statistical procedures for relating ratings and learning, or deficiencies in all of these. Sorting out and correcting for these deficiencies will not be an easy task.

Thus, it seems to be that various schemes of ratings by colleagues usually yield discouragingly low reliabilities, and the question of validity has scarcely been touched. Therefore, some scholars (3) (5) (7) (9) have maintained that such ratings really should not be used at all for summative evaluation. Their reasoning is that error is just about as likely to emerge as truth and that that is an unacceptable situation when a person's rank and salary and, perhaps, very career is in the balance.

This point of view has to be taken very seriously but, if accepted, it leaves us with a serious dilemma. I think it says that systematic evaluation of teaching effectiveness cannot become a part of the faculty

evaluation procedure until such time as procedures can be developed which are psychometrically much improved. But then, what do we do in the meanwhile? We can use informal impressions of teaching effectiveness, as is sometimes done now, but surely informal impressions cannot be more psychometrically acceptable than a systematic rating process. This seems to me only to say that you can be as unfair as you want, providing you do not have the clear evidence of it. Another option, of course, is that we can do as so many of our institutions do, which is—for all practical purposes—to ignore teaching altogether in the evaluation process. In using scholarship as virtually the only criterion, at least we have a well-understood system in which external judges have made the essential decisions by accepting publications in journals of greater or less prestige, granting fellowships, consultantships, and other awards. It is this system which at so many institutions helps create the impression that teaching is an activity of low priority. But what do we know about the reliability or validity of these judgments? Is there sufficient reason to believe that, except for our being accustomed to them, they are substantially better than systematic judgments about teaching? Perhaps it is to avoid all such problems that some colleges and universities and, I believe, most public elementary and secondary schools base all their salary and retention decisions upon degrees earned and years of service. At least these data will be highly reliable.

I am not ready to recommend that. Instead, I suggest that many kinds of information be channeled into the administrative decision-making process and let the individual or committee which has the decision to make sort out the consistencies and inconsistencies and make judgments about which data are likely to be better than which other data. I believe that conscientious faculty, working toward consensus, using many sources of data can achieve acceptable results. I further believe that collected peer data, fallible as it is, has a place in that process and will in fact improve upon it. Meanwhile, if we can encourage the use of information of this kind, it will give us the opportunity to try different approaches, study the results, and one would hope, gradually improve them.

Let me conclude by saying that I see the whole issue of peer evaluation of teaching to be in a very fluid state. It seems fairly clear that a great deal more experimentation and development must go on to find means of involving faculty which are practical in view of other interests faculty have and in ways that faculty will see as fair and effective.

Present pressures for instructional improvement and the evaluation of teaching may not continue. They may be just a phase or fad which education is prone to and which may disappear having had little effect. Or faculty unions and governmental monitoring agencies may opt for reward procedures which simply use measures easily countable such as years in service, students taught, or hours of post-graduate education.

But if the present discernible, but still weak, trend toward a renewed emphasis on instructional improvement and the evaluation of teaching continue, then we will have not only the opportunity but indeed the responsibility to develop practical and credible methods of involving peers in the procedure.

BIBLIOGRAPHY

1. Bergquist, W. H. and Phillips, S. R. A Handbook for Faculty Development. Washington, D.C.: The Council for the Advancement of Small Colleges, 1975.
2. Centra, J. A. Strategies for Improving College Teaching. Prepared by ERIC Clearinghouse, Washington, D.C.: American Association for Higher Education, December, 1972.
3. Centra, J. A. "Colleagues as Raters of Classroom Instruction," Journal of Higher Education, May/June, 1975, pp. 327-333.
4. Doyle, K. O., Jr. Student Evaluation of Instruction. Lexington, MA: Lexington Books, 1975.
5. Doyle, K. O., Jr. and Crichton, L. I. "Student, Peer and Self-Evaluation of College Instruction." Unpublished manuscript, 1978.
6. Eble, K. E. The Recognition and Evaluation of Teaching. Washington, D.C.: American Association of University Professors, 1971.
7. French-Lazovik, Grace. Evaluation of College Teaching Guidelines for Summative and Formative Procedures. Washington, D.C.: Association of American Colleges, undated.
8. Gage, N. L. "Student Ratings of College Teaching: Their Justification and Proper Use." In Allen, Melnik, and Peele (Eds.), Reform Renewal Reward. Amherst, MA: Clinic to Improve University Teaching, University of Massachusetts, 1975, pp. 121-135.
9. Kulik, J. A. and McKeachie, W. J. "The Evaluation of Teachers in Higher Education." In F. N. Kerlinger (Ed.), Review of Research in Education, Vol. 3, Itaska, IL: F. E. Peacock Publishers, Inc., 1975.
10. McIntyre, Charles J., Seibert, Warren F. and Owens, Richard E. Criteria VI, Evaluation of College Teachers. University of Michigan: Center for Research on Learning and Teaching, Fund for the Improvement of Postsecondary Education, Department of Health, Education and Welfare, May, 1977.
11. University of Iowa. "Proceedings, Peer Evaluation Conference." Unpublished manuscript, 1978.